August 1984

DCIEM NO. 84-R-23

AUDITORY DISPLAY OF PASSIVE SONAR INFORMATION

Sharon M. McFadden

Defence and Civil Institute of Environmental Medicine
1133 Sheppard Avenue West, P.O. Box 2000
Downsview, Ontario M3M 3B9

DEPARTMENT OF NATIONAL DEFENCE - CANADA

Table of Contents

ABSTRACT 0
INTRODUCTION 1
CAPABILITIES OF THE EAR 1
- Detection 2
- Simple Discrimination 4
- Binaural Listening 5
- Relevance to Aural Displays 7

FEATURE ANALYSIS 8
- Peripheral Processing 8
- Feature Extraction 9
- Classification 12
- Relevance to Aural Displays and Training 14

TOP-DOWN PROCESSING 15
- Recognition of Different Sources 15
- Speech Recognition 16
- Recognition of Nonspeech Sounds 18
- Relevance to Aural Displays and Training 19

TRAINING 20
- Whole vs Part Methods of Training 21
- Prompting Versus Feedback 23
- Verbal Descriptions 24
- Realism 24
- Computer Aided Instruction 26
- Relevance to Training 27

ENHANCING AUDITORY DISPLAYS 28
- Bisensory Presentation 29
- Automatic Alert Systems 30
- Signal Modification 31
- Cueing 32
- Relevance to Aural Displays 33

CONCLUSIONS 35
- The Current State 35
- What needs to be Done? 35
- Conclusion 37

REFERENCES 39

ABSTRACT

Although the value of aural displays is generally acknowledged, their use in modern sonar systems is limited because of the requirement for extensive operator training and experience before the systems can be used effectively. For aural displays to be useful in future sonar systems, better methods of presenting information aurally, and more effective training methods must be developed. This report reviews basic and applied research relevant to the development of enhanced aural displays and better training methods. It is not intended to provide precise answers on the value of aural displays or the form they should take; rather, the report is a working paper for those involved in providing such answers. Relevant basic research in psychoacoustics, pattern recognition, and cognitive processing is reviewed for possible innovative methods for improving aural displays. Previous applied research is evaluated to alert the researcher to what methods have been tried and to the many pitfalls in effectively assessing training procedures and signal enhancement methods.

INTRODUCTION

In the early days of passive sonar, sounds picked up by hydrophones were presented only to the ear. The operator, listening to these sounds, decided if they were interesting, classified the interesting ones, and tracked them. It was a difficult task, requiring experienced personnel, and was almost impossible in a high noise environment, such as an aircraft. In later systems, visual displays were used, which largely avoided the problems of masking. Using the visual displays, operators could reach reasonable performance levels after much less training. Threfore, in many sonar systems, the aural display is not used or sometimes is not even included. This is unfortunate, because certain information is more readily (and possibly only) interpretable by the ear (Urick, 1975, Smith, 1982). If this aural information is to be used, more effective presentation and training methods must be developed.

Research on the perception and presentation of complex sounds is carried out in several fields. Basic psychoacoustics investigates how the ear detects and discriminates sounds, but tells us little about how they are classified. The recognition and identification of sounds are studied under the headings of pattern recognition, information processing, and cognition. These fields consider the processing of complex sounds, especially the part played by cognitive factors such as context, syntax, and general knowledge. Applied studies that evaluate different displays and training methods provide necessary practical information.

This paper reviews research in psychoacoustics, pattern recognition, presentation methods and training. The first section summarizes the capabilities of the ear for detecting and discriminating sounds. Section two looks at feature analysis or "bottom up" processing, and section three looks at the role of cognitive factors such as context and knowledge, or "top down" processing. Sections four and five cover more practically oriented research into means of improving performance through better training methods and enhancement of displays. The last section summarizes the most promising approaches and suggests directions for future research. No definitive answers are provided on the value of aural displays or what form they should take. This paper is intended as a starting point for those involved in research to answer such questions.

CAPABILITIES OF THE EAR

Researchers in psychoacoustics have been interested mainly in the detection and discrimination of simple sounds. The studies they have carried out provide useful data on the initial processing of all types of sound and the basic capabilities of the auditory system. This section provides a brief summary of findings on the detection and discrimination of simple sounds, and their relevance to the development of aural sonar displays.

Detection

The sonar operator's initial task is to detect the presence of a target in a noisy background. Although sonar target waveforms are usually much more complex than the signals used in psychoacoustics, psychoacoustic data has proven useful in predicting the inherent detectability of such sounds. A basic knowledge of the detection capability of the ear is necessary to enhancing the operator's capability to detect signals aurally.

Studies in detection measure the ability of the ear to "hear out" a sound (usually called the signal) in the presence of other sounds (usually called maskers or noise). The detectability of a given signal in a given masker is measured by comparing the effective energy in the signal with the effective energy in the noise (signal to noise ratio or SNR). The term effective is used because not all the energy in the signal or the noise may be used by the listener. For stochastic waveforms, the usual measure is the power per unit bandwidth (N_0 or S_0). Detection theory states that detection of a signal by a mathematically ideal observer varies as a function of E/N_0, where E is the total signal energy, S_0t, and t is the signal duration. In real conditions, the detectability of a signal depends on the spectrum of the signal and the masker, the duration of the signal, and the listener's knowledge of the stimulus.

An important concept in signal detection is the "critical band". The critical band has been defined in different ways, but all definitions contain one of two concepts. The first is that the critical band is the band of frequencies that affects the detectability of a signal, and the second is that it is the band of frequencies over which the ear can integrate. Many studies have attempted to measure the width and shape of the critical band (Patterson and Green, 1978). Studies that estimate the critical band by determining when an increase in masker bandwidth does not result in a further decrease in the detectability of a signal usually find that it is about 50 Hz wide below 300 Hz increasing to a sixth octave at frequencies above 500 Hz. If the capacity of the ear to integrate sounds instead of reject them is looked at, somewhat narrower critical bands of around a twelfth of an octave are usually found (Taylor, Smith, and Tytla, 1975; Feth and O'Malley, 1977).

The critical band model is important in predicting the detectability of a masked signal as a function of the frequency spectrum of the signal and of the masker. If the spectra of the masker and the signal differ sufficiently that they fall within different critical band filters, the masker will have little effect on signal detectability. As the centre frequency of the masker approaches that of the signal, energy from the signal and masker fall within the same critical band and detectability drops (Patterson and Green, 1978). Detectability is poorest when the centre frequency of the masker is the same as that of the signal. An exception occurs when the signal and masker are both tones and are close enough in frequency that beats occur. Under these conditions, detection is poorest when the frequencies of the signal and masker are a few tens of hertz apart.

The ear's use of these band-limited filters prevents rejection of unwanted energy at nearby frequencies but allows integration of useful energy to enhance detectability of narrow-band noises and multicomponent signals. Sounds analysed by a single critical band can be completely integrated, that is two tones slightly different in frequency and equal in intensity will be 3 dB more detectable than either tone presented alone. If the components are in separate critical bands, detection is also a function of the number of components or the bandwidth, but the improvement in performance as a function of the number of components is smaller. Green (1958) found that the detectability of a signal composed of two sinusoids (separated by more than a critical band) could be predicted best by assuming that the outputs of several critical bands were linearly combined. For noise signals whose bandwidth exceeds a critical band, Taylor and Forbes (1969) found that detectability increased at a rate of 1.5 dB per octave increase in the bandwidth of the signal.

As with the integration of energy over frequency, the integration of energy over time occurs over a limited range of durations. Green, Birdsall, and Tanner (1957) found that perfect integration occurred for signal durations between 20 and 100 milliseconds. (The signal energy required for detection is unchanged but the signal power required decrease 3 dB per doubling of duration.) Later studies by Watson and Gengel (1969) showed that the range for complete integration was a function of frequency ranging from 20 to 175 milliseconds at 125-250 Hz and from 20 to 60 milliseconds at 3000-4000 Hz. For tones of less than 16-20 milliseconds in duration, the specification of threshold as a function of duration becomes difficult because of the spread of masking. If noise masked tones are used, the total energy required to detect the signal decreases as duration increases (Green *et al.*, 1957). (The signal power required decreases 4.5 dB per doubling of duration.) However, if the signal is an unmasked tone below 250 Hz, the total energy required to detect the signal actually increases as duration increases up to 20 milliseconds because of the spread of energy upwards into a frequency region of greater sensitivity. For signal durations between 100 milliseconds and about three seconds the signal power required to maintain constant detectability decreases 1.5 dB per doubling of duration (Green *et al.*, 1957). For longer signals, detectability is usually a function of signal power.

The amount of information about the signal and masker available to the listener at the time the signal is presented can affect detection. In a laboratory, the listener usually knows the approximate onset time and has some knowledge about the frequency spectrum of the signal. The masker waveform, if it is noise, varies randomly. Reducing the information about the signal by randomly varying one parameter (e.g. frequency) reduces detectability by about three dB under the worst conditions (Green, 1961). On the other hand, by presenting the same noise sample on each trial, the detectability of a tone can be raised 2 to 5 dB (Spiegel and Green, 1982). If it is possible to provide extremely precise information (moment to moment variations in the waveform) about the signal (but not the masker), detectability can be improved 8 to 10 dB (Taylor and Forbes, 1969).

In an operational setting, little information is available about the signal or the masker. Stallard and Leslie (1974) calculated that the signal to noise ratio at which a signal could be detected would be 5.4 dB lower in the typical laboratory experiment than in an operational setting.

Simple Discrimination

Although many factors affect identification of targets it is the basic discrimination capability of the ear that sets the ultimate limit. Before an operator can identify a target uniquely he must be able to discriminate that target from similar sounding targets. A better knowledge of this capability may enable us to transform signals into a form that takes advantage of the ears strengths.

Psychoacoustic studies of discrimination look at the ability of the ear to detect shifts in the frequency or amplitude of a simple signal, and to label these shifts as up or down. The ear is good at such tasks. Under ideal conditions, intensity shifts of 0.5 dB and frequency shifts of 0.1% can be discriminated. With these small frequency shifts, the ear can tell there has been a change, but cannot tell in which direction it went. If a direction has to be assigned to the frequency change, best performance will be closer to 0.25% for frequencies below 4000 Hz (2.5 Hz at 1000 Hz), increasing to 7% at frequencies around 10,000 Hz (Henning, 1966).

Amplitude and frequency discrimination both depend on signal intensity. Discrimination continues to improve at signal to noise ratios greater than those required for essentially perfect detection. A tone is detected almost all the time in wide band noise when it exceeds the noise spectrum level by 15 to 20 dB. Frequency and amplitude discrimination continue to improve slowly as the SNR increases up to 60 dB (Henning, 1967, and Henning and Bielwas, 1967).

Amplitude and frequency discrimination depend on the duration and frequency of the tones being discriminated, as well as on their intensities. Frequency discrimination is constant from about 250 to 3000 Hz (Henning, 1966), and amplitude discrimination from 200 to at least 8000 Hz (Jesteadt, Wier, and Green, 1977). Both frequency and amplitude discrimination fall off with decreasing duration once signal duration falls below a given value. According to Henning (1970), the critical duration is a function of frequency. For amplitude discrimination, the critical durations ranged from 100 milliseconds at 250 Hz to less than 25 milliseconds above 1000 Hz (Henning, 1970, Henning and Psotka, 1969). For frequency discrimination, the critical durations ranged from almost 200 milliseconds at 250 Hz to 50 milliseconds at 1000 Hz, dropping to 25 milliseconds at 4000 Hz (Henning, 1970).

In addition to the above parameters, experience may influence the ability of a subject in frequency discrimination. Most frequency discrimination experiments use musically trained subjects. With untrained subjects, the minimum perceivable frequency shift is usually much larger and more variable. It is not clear if this difference is due to experience or to inherent ability, but probably both components are involved.

As well as discriminating shifts in frequency the ear can hear individually the lower harmonics in a complex tone (Plomp, 1964). This capability seems to be a function of the critical band. The lower harmonics are separated by more than a critical band while higher harmonics are not.

Binaural Listening

Research on binaural listening provides evidence of the auditory systems capability to discriminate small differences in sounds presented to the two ears. Current aural displays ignore this binaural capability. By using the findings of studies on binaural listening, it may be possible to increase the information presented to the auditory system.

Humans are able to determine not only what sounds are in the environment, but from what direction they come. They do this through two completely independent mechanisms. One mechanism depends on comparison of timing and amplitude of sounds at the two ears, the other on reflections from the external ear (the pinna) that occur at each ear independently.

With headphones, internal reflections in the cavity formed by the headphone and the ear canal override any monaural pinna reflection cues to signal direction. Only the interaural timing delay and amplitude differences can be used to suggest direction. Since the cues that normally occur for sounds in any external direction are absent, the brain resolves the dilemma by perceiving the sound as emanating from a location within the head, on an arc connecting the two ears.

The auditory system is as sensitive to amplitude and temporal differences between the ears as it is to amplitude and frequency shifts in tones presented monaurally. Over the frequency range of 200 Hz to 10 kHz, amplitude differences of about 0.5 dB can be detected. Interaural time delays of as little as a few microseconds can be perceived by highly trained observers under the right conditions. Optimum discrimination of temporal shifts occurs with signals of approximately one second in duration whose spectral content is primarily below 1500 Hz. For tones above 1500 Hz the discrimination of interaural time differences is impossible. With broad-band noises and complex sounds, temporal discrimination is a function of the amount of low frequency information available.

Amplitude shifts and interaural delays are perceived as a difference in position of the sound "image" in the head. If precisely the same information is presented to the two ears, a single fused image is heard in the centre of the head. Changing the amplitude or the relative onset time of the sound at one ear causes the position of the image to shift (referred to as lateralization of the image). With sufficient disparity between the two ears the image becomes diffuse and in some cases, splits into two. Through analysis of direct and indirect data on lateralization, it is possible roughly to correlate image position with amplitude ratio and temporal delay (Durlach and Colburn, 1978).

The effect on image position of changing interaural amplitude is generally straightforward. Similar effects are found whether the sound is a high frequency tone, a click, a noise pulse or a pulse train. As the intensity ratio increases from zero to approximately ten dB, the image is perceived as shifting from the centre of the head to the ear with the more intense sound. At least for a noise burst, the results are independent of the overall level and duration of the sound (Pinheiro and Tobin, 1969).

The effects of interaural time delays are much more complex. For tonal signals, the effect of changing interaural delay depends on the frequency of the signal. Above 1500 Hz, the position of the image is not affected by altering the interaural time difference. Below 1500 Hz, interaural phase difference is important, rather than interaural time difference. When the signals in the two ears are in phase, there is a clearly fused central sound image. As the phase difference increases, the fused image moves toward the leading ear, but as the phase difference approaches 180°, the image becomes more and more diffuse and difficult to locate. As the phase shift increases, the perception is equivalent to a lead at the other ear, and the image fuses again at the other side of the head (Sayers, 1964). The amount by which the fused image moves from the centre of the head is a function of frequency: near 1500 Hz there is little movement, no matter what the phase, but at 700 Hz or below, the image can move as far as the ears, or even beyond if the phase shift corresponds to an "unnatural" time delay.

For a continuous broad-band noise, a click, or similar type of waveform, lateralization depends on interaural time delay rather than phase. As the interaural delay increases from zero to one millisecond, the image shifts from the centre of the head to the leading ear. For continuous broad-band noises, as the interaural delay increases beyond one millisecond, the image remains at the leading ear but becomes more diffuse until it fills the whole head (Blodgett, Wilbanks, and Jeffress, 1956). At this point, it is indistinguishable from the image produced by statistically independent noises in the two ears. Increasing interaural delay beyond one millisecond with clicks causes the image to split into two (Babcock and Sutton, 1966). With an interaural delay of about 15 milliseconds, the images are located at the two ears. The figure of one millisecond is associated with the headwidth delay while the figure of 15 milliseconds is thought to reflect interaural "memory" (Durlach and Colburn, 1978).

Binaural presentation of masked targets can lead to substantial improvements in their detectability. the improvements arise out of the lateralization capability of the ear, although the exact relation between lateralization and binaural unmasking is not understood. Binaural unmasking occurs when either the signal or the masker are presented out of phase to one another at the two ears. The improvement in detectability can be as much as 12 dB. Practical application depends on being able to independently manipulate the masker and the signal.

Relevance to Aural Displays

Sonar system designers have used psychoacoustic data mainly for modelling the detection performance of a sonar operator. Andrews and Hovater (1971) and Stallard and Leslie (1974) set up tables that predict at what range and in what part of the spectrum a particular type of target would be detected under given environmental conditions. The tables assist the designers of sonar systems and aural display in comparing performance on their systems against expected performance.

The critical band construct has been used in modelling the processing of complex information. Most models assume that the incoming sounds undergo a preliminary transformation in which the ear is modelled as a filter bank followed by energy detectors. The two transforms most commonly used to simulate the filter bank are the Fourier transform and a third-octave analysis. The third octave analysis is used although the widest estimate of the critical band is around one-sixth octave. Fourier analysis has no justification from psychoacoustic data. Nevertheless, both transforms have proven useful in in analysing speech sounds in automatic speech recognition systems.

Psychoacoustics data may assist in determining under what circumstances aural displays should be used and what form they should take. The capability of the ear to integrate energy from different parts of the spectrum and to detect small, rapid shifts in amplitude and frequency make it useful for detecting broad-band and transient sounds. The most useful visual displays, in contrast, show a narrow portion of the frequency spectrum over a period of several minutes. Such displays are best for perceiving narrow-band stable sounds which the ear is not as effective in processing.

Psychoacoustic data may provide insight into how to enhance the perceptibility of information presented aurally. Optimum amplitude and frequency discrimination occurs between 200 and 3000 Hz (the frequency range for speech). In contrast, many sonar targets radiate below 200 Hz. Hence, shifting the incoming sounds upwards in frequency before presenting them to an operator may be useful. As well, there is little advantage to presenting a stimulus to the ear for a prolonged period of time. The integration period of the ear is less than half a second. Time compression of the typically slow changing sonar sounds would allow more efficient and possibly more effective processing of the incoming information. The rate of change would be closer to that favoured by the ear. A

time compressed aural display would give the operator time to compare the output of several channels or compare the output of one channel with a stored template. A display that allowed rapid scanning of the different channels might be particularly effective for initial detection. Comparison of different channels could also be facilitated by making use of the lateralization capability of the two ears. Adjacent channels could be presented at different locations in the head to allow a more global picture of a particular segment of the ocean.

The main limitation with most of the psychoacoustic research is that it deals with relatively simple, short duration sounds. Sonar sounds are typically complex and long duration. The detectability of a complex sound is assumed to be a function of its most detectable component, while the detectability of a long duration sound is assumed to be a function of its power. However, there has been little direct study of the detectability of such sounds. Other factors such as temporal changes and vigilance may affect detectability in ways that cannot be predicted from existing data. Even less is known about the discriminability of long duration complex sounds. Some basic research using relatively long complex sounds could prove very useful in predicting operator performance more precisely and in developing better aural displays.

FEATURE ANALYSIS

The human uses many different sources of information in classifying a complex sound. Of primary importance is the information available in the sound itself. Considerable effort has been spent on modelling the processing of this perceptual information. One of the most useful constructs in the modelling of perceptual processing has been feature analysis. This section examines feature analysis and its utility in the development of better aural displays and training programs for using aural displays.

Feature analysis states that a sound can be represented by a set of features such as its pitch, periodicity, noisyness etc. Two distinct sounds must differ on at least one feature (eg. the pitch of one sound may be high that of another low). According to feature analysis, recognition involves three stages. The sound is first transformed into a sensory representation and then into a set of features which is compared with the features of known sounds stored in memory.

Peripheral Processing

The first stage, peripheral processing, involves a preliminary analysis or "measurement" of the incoming sound (Howard and Ballas, 1981). The ear carries out a spectral and temporal analysis of the sound during this stage, the result of which is a representation of the original waveform in a frequency-time space.

Descriptions of the peripheral processing stage are based on the results of psychoacoustic research. Thus, the output of the spectral analysis is frequently modelled as the output of a filter bank that approximates the critical bands. Both Plomp (1970) and Howard and Burgy (1978) used a third octave filter as a model of the preliminary analysis stage. Plomp (1970) felt that the frequency-analysing power of the ear agreed well with a third octave filter bank. However, as stated in the previous section, the critical bands are more closely approximated by a sixth octave or possibly even a twelfth octave filter bank. A comparison of the utility of the different models would be useful. Howard (1977) did compare the output of a one Hz filter bank with a total bandwidth of 135 Hz, a one Hz filter bank with a total bandwidth of 1080 Hz and a third octave filter bank in his study modelling the processing of underwater sounds. He concluded that the third octave filter best simulated the frequency-analysing power of the ear.

While the output of a filter bank seems to be a reasonable representation of the steady state information available to the feature analyser, it does not account for the ear's perception of change. As discussed earlier, the ear is good at discriminating subtle and rapid changes in frequency and intensity. To get an accurate representation of the output of the preliminary analysis stage, a model of the temporal analysing capability of the ear is needed. To date, little has been done in this area.

Feature Extraction

After initial processing, the next step, according to most feature analysis models, is to reduce the sound to a set of features that uniquely describe it. If extraction of the features is essential to classification, as the feature analysis model claims, then a good understanding of this stage could be useful for both training and enhancement. By knowing what the underlying features are, we should be able to tailor training and enhancement techniques to extracting those features. By understanding how the extraction process goes on, it may be possible to develop techniques that enhance or even mimic that capability. Research efforts have concentrated on determining whether classification of complex sounds involves feature extraction, what the underlying features are, and how the features are extracted. Most of the research has been indirect. Observers are asked to categorize a group of multidimensional sounds. A category is usually experimentally defined as a set of values (the features) a sound can assumes on a set of dimensions. The sounds in each category have a different set of values. A subject's success in categorizing the sounds presented is taken as evidence that the auditory system transforms a sound into its underlying features.

Most studies find that subjects can categorize a set of sounds on the basis of experimentally defined dimensions. Typically, the dimensions chosen are physical characteristics of the sound. An example is the work of Webster, Carpenter and Woodhead (1968a, 1968b). They asked observers to classify a sound on the basis of its relative amplitudes of the harmonics (timbre). The subjects

learned to identify tones with fundamentals of 98, 101, and 121 Hz. Later they were tested on tones with fundamentals of 75 and 150 Hz. Performance was similar to that found with the original tones. They concluded that the subjects were identifying the sounds from the underlying structure (the amplitude pattern of the harmonics) instead of memorizing particular sounds.

In a later study by the same group (Webster, Woodhead, and Carpenter, 1969), observers were asked to classify engine sounds that differed in engine type and speed. There were nine sounds in all, made by three different engines running at three different speeds. The different engines had different high frequency components whereas different speeds meant differences in rhythmic pattern. Observers learned to identify the different engines operating at different speeds correctly about 60% of the time under optimum conditions. Engine type was identified better than engine speed. In a third study (Webster, Woodhead, and Carpenter, 1973), subjects identified sounds that differed in four dimensions: fundamental frequency - 90 or 142; driving waveform - square or sawtooth; number of formants - 1 or 2; and formant position - 600 (600 and 1550) or 940 (940 and 2440). Once again, performance was consistent with the idea that subjects reduce the sounds to underlying features. Sounds that differed only along the formant dimensions were more often confused then sounds that differed along the frequency and waveform dimensions. If subjects had been memorizing the individual sounds that bias would not have been expected.

Having determined that a subject's performance is consistent with feature analysis, what underlying features does a listener naturally use? Multidimensional scaling has been found useful for specifying the underlying features of a group of sounds. Subjects are asked to make similarity judgements on all possible pairs of sounds in a stimulus set. The similarity measures are subjected to a multidimensional scaling analysis that depicts the data in an n-dimensional space. These dimensions are assumed to reflect the underlying perceptual dimensions of the set of sounds. Howard and Silverman (1975) used this method to identify the underlying dimensions of the 16 sounds used in the identification experiment of Webster *et al.* (1973). Three components were found - one corresponding to fundamental frequency, one to the harmonic amplitude spectrum and a third to a combination of the formant dimensions. Their findings correlated well with the results of Webster *et al.* (1973). They concluded that a similarity task was a useful means for getting at the underlying features in complex sounds and that these features correlated with physical dimensions of the sound.

The value of multidimensional scaling for determining the underlying features of complex sounds has been substantiated in applied studies using real sonar sounds. Howard (1976) asked listeners to make similarity judgements on a set of steady state sonar sounds. Two dimensions were found. One corresponded to the third octave band-level measurements of the eight signals while the second corresponded to the presence or absence of a low frequency periodicity. Mackie, Wylie, Ridihalgh, Shultz, and Seltzer (1981) required experienced sonar operators to carry out a comparison task on the set of steady state

sonar sounds used by Howard (1976) and on a larger set of sounds that included samples of operationally significant targets. The underlying dimensions for the initial set corresponded to those found by Howard, but the saliency of the individual dimensions differed. In the second part of the study, the positions of the targets in the multidimensional space were compared with classification success. The sounds that had strong projections on some of the dimensions determined by the multidimensional scaling analysis were classified with much greater success than those sounds lacking any strong projections.

None of the above studies deals with the question of how the ear might transform the output of the peripheral processing stage into a a relatively small set of features. Several selection processes have been considered, but they are mainly variants of two concepts: the property list approach and the process-oriented approach (Howard and Ballas, 1978). The property-list approach suggests that the human is equipped with a set of specific feature detecting mechanisms. The auditory feature extractor is seen as a filter-like device that monitors the incoming sounds for particular stimulus properties. Each detector is tuned to look for one particular feature such as loudness or pitch. The process oriented approach suggests that the human has internalized a set of rules and processes to establish what the distinctive features should be in any particular stimulus context. The important features in a complex sound reflect whatever structure exists in the output of the initial processing stage. A more complete discussion of these two views can be found in Howard and Ballas (1978).

Studies with complex sounds indicate that both approaches may be reasonable depending on the feature being extracted. For unidimensional attributes such as loudness a property-list approach seems reasonably efficient. However, for features such as timbre which cannot be described so simply (Plomp, 1970), the process oriented approach seems more efficient. Evidence for specific detector mechanisms comes from neurophysiological investigations of single unit responding in sensory systems (eg. Hubel and Wisel, 1962 investigations of feature detectors in the visual system). Evidence for the process-oriented approach has been more indirect. One method compares classification performance or the results of a multidimensional scaling analysis with a physical analysis of the sound. If the results of the two analyses correspond, then the features extracted from the physical analysis are assumed to correspond to the features extracted by the human.

Howard and Burgy (1978) investigated the utility of a process-oriented approach for determining the features that listeners use in comparing complex sounds differing in timbre. They asked observers to make similarity judgements on a set of 16 sounds that differed in timbre. The data were analysed with the ALSCAL nonmetric scaling program and the resulting dimensions compared with those obtained from a physical analysis of the sounds. For the physical analysis, 1/3-octave band-level measurements were obtained for each sound, and a principal-components analysis was carried out on these spectra. Principal-components analysis may be thought of as a structure preserving transformation that maps stimuli (the sound being processed) from one space (the output of the peripheral processing stage) into another of lower dimensionality (the set of

features). The results were similar. The transformation predicted the nature and salience of the two dimensions found by the scaling procedure.

The approach used by Howard and Burgy (1978) appears to be potentially useful for understanding and analysing complex nonspeech sounds. The algorithms should be tested with broader sets of sounds and with sounds of different degrees of discriminability. Of course, the utility of the algorithm depends to a certain extent on how well the preliminary analysis stage is modelled. For instance, a single 1/3-octave band-level measurement removes temporal variation in the sound. As a result, the principal-components analysis would not pick out temporal features even though the human may use them in differentiating amongst a set of sounds (Howard and Ballas, 1978).

Classification

Having reduced a sound to its underlying components, the listener must classify it. According to most feature analysis models, the output of the feature extraction stage is compared with internal representations of sounds or categories. Under ideal conditions, the output of the feature selection stage should allow precise categorization of a sound. Unfortunately, conditions are rarely ideal. Factors such as masking, the sensory limitations of the aural system, and the listeners previous experience result in variability both in the sound being processed and in the the internal representations with which it is being compared.

An overriding factor that the operator has little control over is masking. A masker is any irrelevant sound that may distort the target being analysed. In its simplest form, masking blots out significant features of the target. Sometimes masking takes a more subtle form. Martin (1978) found that when subjects were asked to identify partially masked targets, all of which had a perceptible but irrelevant (to the identification task) feature (amplitude modulation), the task was more difficult than when this feature was absent. Subjects claimed that it was difficult to ignore the amplitude modulation even when they knew it was irrelevant to the identification task. Previous experience with the target set can influence the effectiveness of a masker as well (Watson and Kelly, 1981). For instance, Webster, Woodhead and Carpenter (1970) found that partially masked speech sounds were more easily identified than were less familiar, partially masked, nonspeech sounds.

The sensory limitations of the ear were discussed in the previous section. If the differences between two targets is less than the ear can discriminate then classification will be impossible. In practice, almost any two unique complex sounds are discriminable if presented one after the other, but presented in isolation they may be confused. As shown in the work of Webster et al (1973), sounds that differ along certain dimensions (eg. in frequency) are less likely to be confused than those that differ along a different dimension (eg in formant). There has been very little systematic investigation of the relative discriminability of sounds along different dimensions when the differences are well above

threshold. As has been found in studies of colour perception, above threshold differences may not be predictable from data on threshold differences.

Masking and sensory limitations largely affect the processing of the current sound. Previous experience (which necessarily involves sensory capability and probably masking) mainly affects the internal representations that the listener compares the current sound with. Howard, Ballas and Burgy (1978) looked at the development of internal representations as part of their effort to model the classification process for complex multidimensional sounds. They hypothesized that in the classification process the listener estimates the perceptual distance between a sound and a set of category prototypes. Initially, the variance associated with each category is large, but with experience, it is reduced, and the listener can more accurately assign a sound to the correct category. They further predicted that if one dimension was more important than another during the learning period, the reduction in variability for that dimension would be proportionately greater. (Listeners would be able to discriminate smaller differences along that dimension.) For example, if a set of sounds vary along two dimensions such as pitch and loudness but must be grouped into categories that differ only in pitch, targets differing only in loudness will be easily confused.

To evaluate their model, Howard *et al.* (1978) had listeners classify 16 sounds that varied along two dimensions, into eight categories. The sounds were white noise that was amplitude modulated by sawtooth waves of varying frequency and attack. Four different modulation frequencies (4, 4.8, 5.6, 6.4 Hz) and four different onsets (43%, 57%, 71%, 86% of period) were used. The sounds were categorized in two ways, corresponding to either four categories of modulation frequency and two categories of onset (43%, 57% versus 71%, 86%) or four categories of onset and two categories of modulation frequency (4, 4.8 versus 5.6, 6.4). An analysis of the confusion data showed that the pattern of confusions corresponded to the predictions of the model. Uncertainty was reduced along both dimensions, but the reduction was greatest for the dimension requiring the greatest number of overt discriminations.

Howard and Ballas (1978) compared the attention allocated by different listeners to each dimension in the above classification task with optimum allocation of attention. They were able to estimate the extent to which performance was based on sensory limitation as compared to other factors such as a listener's processing capacity. They found that most listeners allocated more attention than optimum to the most relevant dimension (the one having four categories) and more attention than optimum to the frequency dimension. The former, they felt was because of limitations in processing capacity, the latter because of sensory limitations. They were interested also in whether allocation of attention depended on the category set used or on the subject. Each subject learned to use each category set. A subject's allocation of attention to the most relevant and least relevant dimension was the same regardless of the category set he or she was using at the time.

Classification is probably the most important, but least understood part of the analysis of complex sounds. Webster *et al.* (1973) found that observers could discriminate among their 16 multidimensional sounds 100% of the time but could classify them only 60% of the time. None of the methods they tried for improving performance had any effect. Research is needed to determine methods for establishing the relative importance of sensory limitation and previous experience in the classification of particular sets of sounds. If the problem is previous experience then appropriate training can be provided, whereas if the problem is sensory, enhancement techniques may be possible to improve discrimination.

Relevance to Aural Displays and Training

The constructs of feature analysis have been widely used in the development of training programs (Mackie, 1974). Trainees learned to recognize different features of complex sonar targets and to associate these features with particular targets. In general, operators were not good classifiers even at the end of the training period. Extensive experience in an operational environment was necessary before an operator using an aural display was able to identify targets with a reasonable degree of success.

Two possible reasons for the inadequacy of such training programs were the use of subjectively determined features and an emphasis on feature extraction over categorization. Usually, selection of the underlying features was based on subjective evaluations by experienced operators and trainers. The features they selected may not have been the ones they used in identifying targets. Now, methods are available for objectively determining the underlying features of complex sounds. Multidimensional scaling and principal components analysis have been used to extract the underlying dimensions in sets of sounds and show where each sound lies along each dimension. The relative salience of the dimensions determined by these methods has been shown to correlate with a listener's ability to identify sounds differing along these dimensions. The analysis of real targets may provide a picture of which dimensions should be emphasized in training and which targets require extra practice.

The research on classification shows the importance of emphasizing categorization as well as feature extraction. Good classification performance requires the development of internalized category prototypes. Training should emphasize the development of categories through practice in associating particular kinds of sounds with certain categories.

Feature analysis research may also provide insight into the development of automatic preprocessors and other operator aids. For instance, structure preserving transforms such as principal-components analysis might be used to analyse incoming targets. The output of the analysis could be compared with stored feature representations of known targets. The resulting data might be a useful source of information for the operator.

Both multidimensional scaling and principal components analysis could be used to develop libraries of features of different classes of targets. An operator could then compare an incoming sound with a recording of a known feature to see if they were similar.

The real usefulness of algorithms such as multidimensional scaling and principal components analysis remains to be determined. The studies by Mackie *et al.* (1981) with operational targets are a step in that direction. Similar studies involving principal components analysis are needed. Studies are needed also on the value of experimentally determined features for training and on the effect of noisy signals on the usefulness of the algorithms. For automatic processing, the algorithm must be able to cope with noisy signals.

TOP-DOWN PROCESSING

Sounds are usually processed in a particular context. A sonar operator is aware not only of the auditory patterns of sounds of different targets but also of such factors as the likelihood of certain ships being in the vicinity, the activity pattern over time of different ships, and the effect of weather on sound transmission and ship behaviour. In addition, he has his own set of personal experiences in classifying targets. All these different kinds of knowledge may influence his identification of a particular target.

The use of knowledge other than the acoustic information available in the incoming sound is frequently called "top-down" processing. Research on top-down processing has ranged from studies of relative simple principles the auditory system appears to use to separate incoming sounds by source, to studies on the effect of immediate and general context on the classification of words in continuous speech. The role of top-down processing has been studied most extensively in the recognition of words in connected speech where both semantic and syntactic factors have been shown to influence perception. More recently, research has been directed towards the effect of knowledge-driven processes on the perception and identification of complex non-speech sounds. This section reviews pertinent research on top-down processing to determine the role it plays in the classification of sonar targets.

Recognition of Different Sources

The sounds being picked up by a passive sonar system are usually radiated by several sources. Before a sonar operator can identify a particular target, he must be able to group together the components from a single source and to segregate information from different sources. The auditory system seems to be good at sorting out input into separate sources. Bregman (1978) postulated that the ability of the auditory system to separate out different sources of information is influenced strongly by Gestalt principles such as similarity, closure, proximity, and good continuation, principles that are not based on the

specific experiences of individuals.

The research of Bregman and his colleagues has concentrated on showing the use of Gestalt principles in the perception of simple sounds. In an early experiment (Bregman and Campbell, 1971), six low frequency tones were played alternately with six high frequency tones. Under these conditions, two melodies were heard, the high frequency tones being perceived as coming from one source and the low frequency tones from another. The perception of two melodies follows the principle of similarity - elements will be grouped together if they are similar (e.g. close together in frequency). A related experiment illustrated the principle of simplicity (Bregman, 1978). Listeners adjusted the rate of alternation of high and low frequency tones. At a low rate of alternation, the sounds appeared to come from one source, at higher rates from two sources. Good continuation was illustrated using glide tones. Bregman and Dannenbring, (1973) found that connecting the sequential tones with a glide reduced the tendency to perceive them as coming from two sources. A continuous shift in frequency usually means a single source.

Not only do such principles impose an order on the incoming sounds, they influence the perception of them. For instance, Bregman and Campbell (1971) found that listeners could determine the ordering of three tones in the six tone sequence only if the three tones were perceived as coming from a single source (i.e. were either all high frequency or low frequency).

Bregman (1978) considered these principles to be a set of rules that the ear applies to organize the incoming sounds. Such rules do not come from the sounds themselves, but from perceived regularities in the world. A better understanding of what the rules are, and of how and when the ear uses them, could improve our knowledge of the processing of complex sounds.

Speech Recognition

Most of the research on the effect of top-down processing on the recognition of auditory stimuli has been done in the area of speech recognition. While not directly relevant to this paper, an examination of this research may provide insight into the use of top-down processing in the perception and recognition of non-speech complex sounds. A complete review is beyond the scope of this paper. Instead, a few examples are presented. Cole and Jakimik (1978) provide a more in-depth review of the role of top-down processing in the recognition of continuous speech.

A feature-analytic approach to speech recognition requires that a sound be identified as a particular word or speech sound group when the pattern of features extracted from the sound matches a stored pattern. In order for identification to be successful, the auditory patterns must be stable. Words spoken in isolation do tend to have stable auditory patterns. Machine recognition systems that compare a feature representation of the spoken word with stored templates of words can identify isolated words accurately as much as

99% of the time (Reddy, 1976).

The success rate of the same systems for recognizing words in connected speech is much poorer. The reasons given for the low rate of identification are the high variability of the auditory patterns of words in connected speech, and the lack of distinct boundaries between words. Humans, however, are able to recognize words in connected speech with a surprising degree of accuracy. They are also able to recognize words spoken by a wide range of speakers and words spoken in high levels of ambient noise. One advantage the human has over the machine using feature analysis is the ability to use syntactic and semantic information.

The role of syntactic and semantic information has been studied extensively in speech processing. For example, Miller and Isard (1963) showed that both syntax and semantics could affect word recognition. They presented subjects with strings of words that were meaningful sentences, grammatical but semantically meaningless sentences, or just unrelated words. All the strings were partially masked by noise. Word recognition was highest for the meaningful sentences and least for the unrelated strings of words. The use of contextual information in speech understanding has been demonstrated in other ways. Marslen-Wilson and Welsh (1978) demonstrated the use of semantic and/or syntactic constraints in close speech shadowing. They asked subjects to shadow (repeat verbatim) connected prose that contained mispronunciations. The subjects more frequently correctly pronounced the mispronounced words that were highly constrained by the previous context than those that were not. An analysis of the shadowing latencies prior to the restoration of the mispronounced word showed no increase in latencies. It was as if the subject never heard the mispronunciations. Cole and Jakimik (1978) used mispronounced words to show that context could affect the rate at which words are recognized. Subjects were asked to listen to a prose passage and to press a buzzer whenever they heard a mispronounced word. Mispronounced words that were constrained in some way by the context of the prose passage were recognized faster than unconstrained words. An effect was found when the relevant context was the preceding word or phrases and even when it was something as general as the title of the story being listened to.

The studies described above are a limited selection of the wide array of research that has been done in this area. The findings support the importance of context in the perception of speech. They also show that context can both enhance and degrade accurate recognition of the information being presented. Listeners hear what they expect to hear and that may not be what was said. The main sources of information in these studies were syntax and semantics. An individual's specific experiences may also affect his or her perception of words, but these are less amenable to study and classification.

Recognition of Nonspeech Sounds

Context has received less attention in the recognition of complex sounds. In most studies on the identification of complex sounds, small sets of isolated sound samples are used. As with small sets of isolated words, recognition of isolated complex nonspeech sounds can be explained reasonably well using a feature analysis model. In the real world, complex sounds are neither few in number nor are they presented in isolation (without a context). Recent research indicates that context plays an important part in the identification of non-speech sounds.

The value of context has been examined indirectly by Watson and his colleagues (Watson and Kelly, 1981). They looked at the importance of stimulus uncertainty on the discrimination of word-length tonal sequences. They found that stimulus uncertainty defined by the size of the set of tonal sequences played an important part in the ability of subjects to "hear out" components of those sequences. Subjects were asked to listen to pairs of tonal sequences that either were identical or were shifted in frequency (or intensity or duration) on a single tone in the sequence (the target tone) and to indicate if a shift had occurred. Stimulus uncertainty was varied by varying the degree to which the contextual components (the non-target tones in the sequence) varied in frequency, intensity and duration from trial to trial and by either fixing or varying the location of the target tone. It was found that when stimulus uncertainty was sufficiently low that the listeners could learn the set of sounds, discrimination performance approached that found when the target tone was presented in isolation. With high stimulus uncertainty, discriminable frequency, intensity, and duration shifts were about 800 Hz, 7 dB, and 35 msec. (The target tone in the pair of tonal sequences had to be at least 800 Hz apart before the listener knew that the sequences were different.) The detectable shifts for the target tones in isolation was around 1-2 Hz, 1-2 dB and 5-10 msec. Clearly, when listeners are able to set up a context within which to listen to a sound, their discrimination performance is significantly better.

A more direct approach to the study of contextual effects was used by Howard and Ballas (1980). They found that both syntactic structure and semantic information could affect the ability of listeners to classify nonspeech transient acoustic patterns. Listeners were asked to classify sequences of short duration tones either as target or as nontarget sequences. For one group, the sequences of tones which could represent a target were generated by a finite state grammar. For the second group, the same tones were used, but the target patterns were selected randomly. In effect, the only way of discriminating target from nontarget was by memorizing the target patterns. The grammatical group could, in principle, learn the syntax and use that information to separate out target from nontarget. Neither group was told how the composition of the targets was determined. Both groups improved over time, but by the end of the test, the grammatical group's performance was superior to that of the nongrammatical group.

In a second experiment (Howard and Ballas, 1980), real world sounds were substituted for the tones. Otherwise, the experiment was unchanged. Once again, the grammatical group did much better but the difference between the two groups was smaller. The experimenters concluded that structured patterns of unrelated familiar sounds might be more difficult to identify then structured patterns of simple tones. In a third experiment, the sounds were semantically related. The sounds all had to do with water - valve opening, water drop, steam, clang, water flush. The ordering of the sounds in the targets was logical for the grammatical group, and random for the nongrammatical group. Half of each group was also given information about the source of the sounds. The performance of the grammatical group was superior to the nongrammatical group's performance and the difference was greater than the differences between the two groups in the previous experiments. The additional information about the source enhanced the performance of the grammatical group and degraded the performance of the non-grammatical group, although not significantly in either case.

The importance of context in the recognition of sonar targets was shown in a study by Smith (1983). An analysis of the classification performance of sonar operator's showed that they formed an hypothesis early in the analysis of a potential target and that later analysis was biased towards confirmation of that hypothesis. In some cases, information was ignored or unrealistic error limits were allowed on calculations in order to confirm the hypothesis. The analysis was driven by the context (the hypothesis) rather than the acoustic information.

The research reported above is only a start. The value of top-down processing on the recognition of non-speech sounds needs more study. While we can show the effect of context on performance and to a limited extent predict the effect of different forms of context, little has been done on using context to enhance recognition as opposed to biasing recognition. Before this is possible, more needs to be known about the relative effectiveness of different types of context and the interaction of top-down and bottom-up processing.

Relevance to Aural Displays and Training

An experienced sonar operator is much better than a novice at using aural displays. On the basis of bottom up theories, one might conclude that the experienced operator had internalized more precise templates. Research on top-down processing suggests additional reasons. Context has been shown to play an important part in the recognition of complex sonar targets (Smith, 1983). The experienced operator has a broader range of experience and knowledge than the new operator. With this information he can form a more precise picture of what targets are likely in a particular context. The inexperienced operator knows that A leads to B. The experienced operator also knows that A leads to B but only if C is present and D is not.

Since an operator uses his existing knowledge base it should be made as broad and as relevant as possible during training and under operational conditions. During training, experience should be given with a wide range of targets. The same target should be presented under differing conditions and the differences between similar sounding targets should be emphasized. The value and place of outside sources of information should also be stressed. Under operational conditions, a similar policy should be followed. Emphasis should be placed on the integration of information from different sources and the dissemination of such information to sensor operators. In addition, when faced with certain acoustic information, operators should be made aware of what targets might have those characteristics. A computer aided classification system that accomplishes the latter was described by Smith (1983). The operator enters acoustic information into a computer containing a database of potential targets and the acoustic features that describe them. The computer searches its data base and prints out a list of possible targets having those acoustic characteristics. Properly used, the system encourages an operator to consider alternative hypotheses that he might not otherwise think of. The system works best when the target information entered into the computer is sufficient that a list of four to five possible targets is output.

There are two problems that must be kept in mind in any effort to more effectively use context. The first problem is the limited knowledge available about what information an operator is actually using in making a decision and the weight he applies to different pieces of data. Secondly, it is rare that an operator has sufficient information to completely describe a target. Thus, while it may be possible to provide an operator with alternative hypotheses, the operator may not be able to obtain enough additional information to eliminate some of those hypotheses. Future research should be aimed at dealing with these problems.

TRAINING

Research on auditory perception and cognitive processing can provide only partial answers to the question of how to improve identification of complex sounds. Most laboratory studies use trained, highly motivated observers and small, simple sets of sounds. To use the results in an operational setting, applied studies are needed to confirm the laboratory findings.

This section reviews some of the research on the effect of different training methods on the learning of complex nonspeech sounds. A brief examination of the potential role of computer aided instruction is included. Most training methods studied have proven ineffective. There appears to be no easy way to get around the need for extensive experience with targets.

Whole vs Part Methods of Training

Methods of teaching the recognition and identification of complex stimuli are usually divided into two major types. With the first type, one or more of the stimuli to be identified are presented to the observer on successive trials. This is known as the whole, or synthetic, method because the emphasis is on perceiving the whole stimulus. With the second type, the part or analytic method, the subject is taught to recognize the features of which the stimulus is composed. For instance, in an aircraft recognition task, with the whole method the students are asked to identify examples of different types of aircraft. The part method, on the other hand, emphasizes the recognition of the different parts of the aircraft (e.g. wings, fuselage) followed by the association of particular features with different types of aircraft. The relative usefulness of these two methods has not been definitively determined.

An extensive study of the part method was carried out by Annett (1971). It illustrates a useful approach to studying different training methods and the pitfalls that are often encountered. Annett contrasted two types of part methods, called part/synthetic and part/analytic, with a simple whole method. The task was to classify each target as being one of four different types of ship. Each ship was characterized by several features but a target's category could be identified uniquely by listening to one feature — the propeller cavitation rate. In the whole method, a target was presented on each trial and the subject was required to label it. In the part/synthetic method, the subjects first learned to identify the different sounds that the targets could make, and then they learned what sounds were associated with each target. Finally, they were tested with the whole target. There were two versions - one in which all possible relevant features (experimenter determined) were taught, and a second in which only one feature (that took on a different value for each class of target) was taught. With the part/analytic method, the whole sound was always played, but during the training trials, one feature was played louder than the rest for part of the presentation period. The augmented feature was either experimenter determined or subject determined. The "part" group that learned initially to identify propeller cavitation rate and the "whole" group performed similarly and their performance was superior to other "part" groups that were exposed to multiple acoustic features.

These results are not surprising. The "whole" group presumably learned to use propeller cavitation rate to discriminate the sounds. The other "part" groups were hindered because they had less experience with the relevant dimension. To compare the various methods adequately, it should not have been possible to identify a target by listening for a single feature.

A study comparing an analytic and synthetic method but using sounds that did not all differ along the same dimension was carried outby Swets, Millman, Fletcher, and Green (1962). In their analytic method, subjects identified the sounds by a number. Each digit in the number corresponded to the value the sound had on a particular dimension. For example if the first digit stood for frequency and the sounds could be at one of five frequencies then the

first digit would be between one and five. In the synthetic method, the sounds were identified by an arbitrary name. They concluded that their analytic method was superior to their synthetic method. Unfortunately, the same sounds were not used in the two methods.

A somewhat different approach to studying the relative usefulness of part and whole training methods was reported by Mackie (1974). He described two studies on the usefulness of the part method. In the first study, performance was compared across three different training methods. In two of the methods, the subjects first learned to identify the specific features of the targets and then to identify a target using the clues they had learned. Both groups were given information on what features were associated with different targets. The groups differed in that one was trained to use a formal decision logic. The third method followed the standard format of presenting the target and having the subject label it. All groups scored around 75% on identifying the presence or absence of eight different clues in target sounds. On the target identification task, the part group using the formal logic identified about 10% more targets than the other two groups.

The second study reported by Mackie (1974) evaluated the performance of experienced operators as a function of the information available to them. For each sound, the operator went through four steps. In step one, he listened to the sound for ten seconds and then classified it. Next, the sound was replayed for five minutes and the operator analysed it and reclassified it. In the third step, the operator judged the presence or absence of eight clues and took a beat count. A third classification was then made. Finally, a fourth classification was made using the correct answers to the questions asked in step three. Across the four steps the numbers of correct classifications were 39%, 44%, 49% and 57%. Since there was no control group it is not possible to say how much of the improvement was due simply to continuing exposure to the targets.

The above studies do not provide a clear cut answer about whether a whole or part method is best. They do show different methods that can be used to study this question, and some of the many factors to consider in designing a study.

In the operational world, a part method is usually used for training. The general finding is that initial training does not lead to high levels of performance and that experience is the only real teacher. Of course, it is impossible to say what experience teaches. It could be experience with the target, learning to associate features with the target, or context.

Probably, both part and whole methods play a part in the learning of complex sounds. For learning to read, which is a visual task of similar character, Taylor and Taylor (1983) have suggested a three-phase learning sequence that incorporated both whole and part learning. Phase one involves whole pattern learning of up to a few tens of words. The second phase involves learning to extract common elements from various words. Phase three is learning to put

the elements together as whole complexes. Such a model might be a useful approach to training operators to identify complex sounds.

Prompting Versus Feedback

In training, a listener learns to associate a sound with a label. The label may be presented before, during, or after the sound. If it is presented before, it is frequently known as a prompt. If it is presented after the sound has stopped and a response has been made, it is called feedback. A label presented while the sound is on could be either a prompt or feedback depending on whether it is presented before or after the response. Most researchers agree that it is important that subjects have a label to attach to the sound they are hearing or have just heard (Annett, 1971). The importance of the timing of the label relative to the sound presentation is less clear.

Swets *et al.* (1962) examined the relative usefulness of prompting and feedback in their study on the learning of complex sounds. They compared one prompting and three feedback methods. The prompt was presented before and while the sound was played. No overt response was required. The three feedback methods required the subject to label the sound. After labelling the sound, the subjects in the first feedback group saw the correct label, the second group saw the label and heard the sound again, whereas the third group saw the label and heard the original sound plus the sound associated with their response. The prompt method was superior to all the feedback methods. They concluded that the prompt method interfered least with the association formed between the label and the sound.

Sidley, Winograd, and Bedarf (1965a, 1965b) thought that the finding of Swets *et al.* was due to the concurrent presentation of the label and the sound. In their initial study (1965a), they compared two feedback methods. In both, subjects were required to label the sound before feedback was given. In one, the sound continued during the feedback stage whereas in the second, the label was presented one second after the sound had stop playing. They found that presenting the sound concurrently with the label was superior to presenting it afterwards. A second study (1965b) examined the importance of requiring a response. With temporal overlap controlled, they found no difference between requiring and not requiring an overt response.

According to these studies, there is no advantage to requiring an overt response, but there is some advantage to overlapping the presentation of the target and the label. It must be remembered that these are laboratory studies using highly motivated alert listeners. The results might be different in an operational setting. Requiring an overt response may keep trainees alert. On the other hand, it might interfere with the association process.

Verbal Descriptions

Sonar operators are normally provided with verbal descriptions of different targets, to assist them in learning to identify the sounds. The value of such descriptions seems intuitively obvious, but the research on verbal descriptions suggests that they may be worthless unless considerable care is taken.

Corcoran, Carpenter, Webster, and Woodhead (1968) attempted to find the optimum form for verbal descriptions. They compared the relative usefulness of (1) simply labelling the sound, (2) giving the listener a description of its acoustical characteristics, (3) presenting the label for five sessions then adding the verbal descriptions, and (4) giving the listeners descriptions generated by a previous group of listeners. Subjects were asked to classify 104 targets into 14 different categories. Each sound was defined by the value it assumed on between two and four dimensions. The group that was given the verbal descriptions right from the start learned faster and achieved a higher level of performance, than the other three groups. The group using the experienced listeners' descriptions did best for the first few sessions, but never improved further.

Annett (1971), using sounds that could be discriminated along a single dimension, found verbal descriptions of the acoustical characteristics of sounds unhelpful. Possibly, the necessary information could be retained in the head, so that there was no advantage to providing a written version. Woodhead, Carpenter, and Webster (1973) also found little advantage to giving the listener descriptions of the aural characteristics. In their study, 3 of the sounds were buzzes, three were musical tones and three were vowels. The verbal descriptions enhanced only the identification of the vowels. Another possible explanation is that the descriptions did not provide the listener with useful information. In an earlier study, (Webster, Woodhead, and Carpenter, 1969), they found that providing a meaningful name rather than a code name enhanced learning when the targets to be identified were different types of diesel engines operating at different speeds.

The above results suggest that the value of verbal descriptions depends on the complexity of the sounds and the relevance of the description. Annett (1971) and Woodhead *et al.*. (1973) both used simple targets and found no advantage in providing verbal descriptions. Corcoran *et al.* (1968) demonstrated that the content of the descriptions was important for more complex sounds.

Realism

In the operational world, the targets to be identified are usually partially masked by ambient noise or by other sound sources. If masking modifies the signature, then the trainee should have experience with noisy targets. On the other hand, the research of Howard, Ballas, and Burgy (1978) suggests that classification performance depends on the precision of the internal representation

of the targets. Masking increases the variability of the sounds, and may impair the development of precise internal categories.

Corcoran *et al.* (1968) looked at the identification of noisy targets as a function of experience with noise free and noisy targets. They compared three conditions. The first group was trained and tested with noisy targets. The second group was trained using targets that became progressively more difficult (the S/N ratio decreased) over the first five sessions, but were tested on noisy targets. For the third group, the S/N ratio of the targets was randomly varied in both the training and test portions of the first five sessions. All three groups were trained and tested on noisy targets in a second set of five sessions. The results supported initial training with noise free targets followed by experience with noisy targets. The second group learned faster and better than the first group in the first five sessions and performed best on the last five sessions. The third group performed the same as the first group on the last five sessions. Their performance on the first five session is not comparable since their tests were easier.

Mackie and Harabedian (1964), investigated the usefulness of (1) simple synthetic, (2) variable synthetic, (3) noisy and variable synthetic and (4) real targets for training sonar operators. Each subject was trained using one type of target only. All were tested on real sea recordings. Detection, tracking, and active and passive classification performance were all studied, but only the passive classification study used an aural presentation mode. They found that for passive classification, using sea recordings during training was best. However, they did find positive transfer effects with variable synthetic targets. The addition of noise was detrimental, although it was useful in the detection training. They felt that the use of noise during passive classification training prevented observers from learning some of the more subtle classification clues.

The type of noise as well as the amount of masking may affect learning. Webster, Woodhead, and Carpenter (1970) showed that the spectral characteristics of the masker could affect classification performance even with relatively simple sounds. They compared the identification of three classes of sounds: (1) vowel-like, (2) music-like, and (3) meaningless (buzzes) masked by -6 dB per octave thermal noise or by a harmonic tone having the same fundamental as the targets. In both cases, the masker was set 6 dB below the masked threshold of the target. Identification of the vowel-like and the music-like sounds was poorer when they were masked by the noise, whereas the effect of the two maskers on the meaningless targets was similar.

In selecting the sequence of sounds to be presented during a training program, the best procedure seems to be:

1. present clean realistic targets;

2. present the same targets partially masked by increasingly realistic noise;
3. add in other confusing sound sources.

Computer Aided Instruction

There has been considerable research on computer aided instruction (CAI). Most of it centres on the use of CAI for cognitive learning, whereas identification of complex sounds is a perceptual learning task. For perceptual learning, computers are often used to vary the rate and type of material according to the student's progress. For instance, a computer can be programmed to assess performance, to give additional experience with difficult stimuli, to provide appropriate feedback, and to permit the student to select the stimuli and/or stimulus features for a particular trial or sequence. Two examples of research on the usefulness of CAI are outlined below.

Swets *et al.* (1962) examined the usefulness of CAI for teaching the identification of complex nonspeech sounds. They used the computer to present individually appropriate feedback, to provide extra practice on more difficult items, and to order the presentations of items to emphasize easy discriminations initially and more difficult discriminations later. The sounds to be identified were complex tones. Each tone assumed one of "n" values along each of "m" dimensions. The values of "m" and "n" varied across the set of experiments. The first experiment examined the effect of using the computer to tailor the feedback given to a subject. In the control condition, the subject made no overt response. A label was presented visually and then the sound was played. In the computer assisted conditions, a sound was played, the subjects typed in a response, and then the correct response was typed. Under two of these conditions, the subject was given a second chance to identify the sound correctly. When classification performance was tested at the end of the training period, the control group received the highest scores.

A second experiment examined the effect of using the computer to order the relative discriminability of successive sounds, and to give extra practice on items that were incorrectly identified. In the control condition, each successive sound was selected at random without replacement from the total set. The second condition replaced those items that the subject identified incorrectly. The third condition was like the second except that successive items were not randomly selected. A pair of items were chosen and presented alternately until both were identified correctly. Members of successive pairs became increasingly similar until the last pair differed by a single value on a single dimension. In all conditions, when the set had been exhausted, the entire set was reinstated. The first condition, which provided equal experience with all the sounds, resulted in the best performance.

Swets *et al.* concluded that the special techniques made possible by the computer not only failed to improve the level and rate of learning, but impaired learning. Their subjects learned best under the methods that did not require a computer. However, they felt that their findings might not be generalizable to the real world. In an experiment, subjects are highly motivated, and left to their own devices will usually perform optimally. In a real training environment, students are not as highly motivated, and the CAI techniques may be useful in sustaining performance over a prolonged learning period.

Mackie (1974) examined CAI in an operational environment. He was interested in finding a better method for teaching sonar operators to classify aural targets. The study reported by Mackie did not compare CAI with traditional methods, but it did illustrate that CAI was potentially useful for training sonar operators. The computer was used both for initial and for remedial training. After 3 or 4 hours training, naive subjects could do as well as experienced subjects on clue identification. In the remedial training study, the computer assessed the operators' strengths and weaknesses and then gave them refresher training on the types of targets giving them the most problems. After one day refresher training, substantial improvements were found. Performance on the posttest was a function of the number of items and the kinds of items that the operators had studied during the training period. Thus, recognition of high threat targets, with which the students were less familiar initially, improved substantially, but at the expense of recognition of other types of targets.

Since the study reported by Mackie (1974) used only the automated procedures, it is not possible to say what performance might have been achieved with traditional methods. The automated method did at least permit rapid learning, at a manpower cost much lower than the traditional method.

CAI seems to have a place in the training of sonar operators. Operators can be given extensive experience with a wide variety of targets, especially low probability high-threat targets. Such experience has been shown to be a prime requirement for learning to classify aural targets. The problem is to develop programs that enhance instead of retard learning. As the study by Swets *et al.* (1962) shows, it is difficult to assess the value of different CAI techniques in an laboratory setting. Nevertheless, with careful effort, CAI should be a useful means of improving the capability of sonar operator to use aural displays. At the very least, CAI can provide substantial cost reductions in training students to perform at an acceptable level.

Relevance to Training

The quality and results of studies investigating training techniques is disappointing. Better research is needed if training programs are to be significantly improved. The existing results do suggest some preferable techniques for teaching the classification of nonspeech complex sounds.

- The listener should be given relevant information about the the sounds to be learned.
- Practice should be given on the individual components of a target, the target as a whole, and how to relate the underlying features to the target.
- During training, the label should be presented at the same time as the sound.
- Practice should initially be provided with clear targets, followed by targets partially masked by ambient noise and other sound sources.

It is not known to what extent these techniques are currently used. In any case, their use is not likely to alleviate the need for extensive ongoing training. At best, they will make that training more effective.

Currently, the only real solution to the training problem seems to be the use of automated instruction. Automated instruction systems can be used on ship or shore without supervision. The operator gets more practice with the targets at a lower cost of both time and money. To be effective, effort must be expended to provide a wide range of targets, good analyses of the underlying structures, and useful descriptions of each target and its components.

Few of the training studies reviewed in this section take into account the most recent research on feature analysis and top-down processing. As discussed above, procedures exist for extracting features objectively from real targets. Studies using objectively determined features should provide data on the usefulness of such analysis techniques in the real world as well as providing information on the value of different training methods.

The research on top-down processing supports the value of verbal descriptions. It may also be useful in determining what kinds of information could be most helpful. For speech processing, syntax and context are important sources of information. During training, neither of these sources is available to the trainee.

ENHANCING AUDITORY DISPLAYS

Early sonar systems were simple, with a minimal capacity for processing the incoming information. Today, with the advent of computers and microprocessors, incoming data can be preprocessed and transformed in a variety of ways. The development of useful preprocessing techniques could lead to considerable improvement in operator performance. The problem is to find methods that enhance instead of limit an operator's capabilities. Too often, preprocessing seems to reflect the capability of the machine more than the needs of the human. This section looks at some methods of data presentation that have been tried in the past. The methods include the presentation of information to

more than one sense, the use of automatic alerting systems, modification of the sounds, and the use of target samples or cues.

Bisensory Presentation

One way of improving performance might be to present information to more than one sense. Alternate sources of information give the operator the opportunity to cross check his decisions, since the senses process information differently. Research has shown that, under the right set of conditions, detection and identification of targets can be better when the information is presented to more than one sense modality.

The concurrent visual and aural presentation of complex signals was studied by Halpern (1971). He concentrated on determining some of the conditions in which a bisensory presentation would be superior to a unisensory presentation of the same information. He also varied the type of information presented visually, and duration. Two different visual displays were used. For both, the information was the changes in the intensities of the auditory signal measured at the outputs of a bank of sixteen auditory filters. For the first display (the redundant condition), the filter bank was based on a model of the analog cochlea. The effect was an attenuation of the output at high and low frequencies. With the second display (nonredundant), the output was not attenuated after being filtered so that more information (in theory) was available at the low and high frequencies than was available aurally. The sounds were presented for either 2.5 or 5 seconds. Performance in the bisensory conditions was consistently superior to the unimodal conditions except for the 2.5 second, bimodal, redundant condition where performance equalled the 2.5 second, aural mode. After training, subjects identified 65% of the targets in the unimodal aural condition, the unimodal nonredundant visual condition, and the bisensory redundant condition, and over 90% of the targets in the bisensory nonredundant condition. In a second set of experiments, a time-history visual display was used. Subjects identified 75% of the targets with the visual display, 65% with the aural display and over 95% with the bisensory display.

In general, Halpern (1971) found that bisensory presentation was best when different information was presented to each sense and when the user had time to integrate the information from both displays. If time was limited, the subjects used the display they believed was most informative. For instance, subjects did not appear to use the visual display in the 2.5 second bimodal redundant condition. Performance in that condition equalled the unimodal aural condition.

The value of a bisensory display for detection was examined by Loveless, Brebner, and Hamilton (1970). They asked subjects to detect a signal masked by noise, when presented aurally, visually (on an oscilloscope) or both. The results showed that a bisensory presentation could enhance the detectability of signals. In their study, signal duration was one second, and the information was redundant. These are the conditions that Halpern found least favourable

for bisensory presentation in his identification task. Performance was subject-dependent. If a subject reported a bias towards one or other display, performance with the bisensory display corresponded to that found with the favoured display.

Given the above results, the use of both aural and visual displays for processing sonar returns should be encouraged. In current systems with both types of display, the information is usually nonredundant. The visual display emphasizes narrow, stable spectral lines while the ear picks up broadband noises and transients. Since the duration of targets is reasonably long, the operator has time to compare displays.

Automatic Alert Systems

An automatic alert system analyses the incoming data, and signals the operator when it thinks it has located a threat target. The operator examines the same data and either accepts or rejects the computer's decision. In the past, the value of the automatic alert system has been questioned, because the algorithms used to analyse the data were considerably less discriminating than the human operator. With most systems, either the operator was constantly handling false alarms or the system missed real targets. Secondly, the information load was usually low enough that the operator could process all the data received even without using the automatic alert. With the new complex multi-beam systems, the operator cannot monitor all the incoming data effectively. Computer analysis is clearly necessary. Moreover, improved algorithms should make the computer's reports more reliable. The main remaining question is whether an operator can use the computer's output effectively. Studies on computer assisted detection and identification show that the operator can use the information available from the computer, but not optimally.

The value of computer assistance in a vigilance detection task was studied by Ingleby, Hamilton, and Copeman (1971) and Hartley, Olsson, and Ingleby (1972). Ingleby *et al.* (1971) examined the usefulness of a 75% reliable computer as a function of signal probability (0.1 or 0.5) and timing of the computer assistance (before or after the presentation interval). Subjects were able to combine the information from the computer with the evidence from their ears or their eyes but not optimally. The computer was most useful when the signal probability was 0.1 and the alert preceded the presentation interval, and when the signal probability was 0.5 and the alert followed the presentation interval. Under these conditions, the subjects ability to discriminate signal from no signal increased with computer assistance. Interpretation of the results was hampered because of limitations in the method. Each session was composed of 80 computer assisted trials followed by 80 unassisted trials followed by 80 computer assisted trials. Thus, the differences found could have been because of an improvement in performance on the computer assisted trials or a decrement in performance on the unassisted trials. To get around this problem, Hartley *et al.* (1972) used different subjects in the assisted and unassisted conditions. In their study, the alert was always presented before and during the presentation period for the

signal. They found that computer assistance increased the readiness of a subject to report a target and the confidence with which a target was reported, but did not increase the subject's ability to discriminate a target from noise.

Corcoran, Dennett, and Carpenter (1970, 1972) studied the effect of providing computer assistance on classification performance. In their studies, the computer display showed the probability that the sound had been generated by each of the potential sources. In each session, performance was measured without the computer analysis, then with it, and then again without it. The variables were signal to noise ratio and the reliability of the computer analysis. In all conditions, subjects' performance with the computer analysis was equal or superior to their performance without it. The computer analysis was most helpful when the probability of the computer identifying the target correctly and the probability of the listener identifying the target correctly without the computer were both about 50%. When the probability of the computer correctly identifying the signal was close to 100%, the subjects' performance in the sets without the computer did not improve across sessions. The listeners appeared to ignore the aural input.

These studies suggest that operators can learn to integrate evidence from the computer with the evidence from their senses. However, optimum performance is not easily achieved. If the computer is seen as accurate, the aural input is likely to be ignored; if the computer is seen as less reliable than the operators' own senses, its evidence is likely to be ignored.

The studies on computer assistance have used simple detection or identification tasks, in which the subject processed the incoming information at about the same time as the computer. In the operational world, the operator is not likely to process the incoming information until after he has received an alert, because it is impossible for him to deal with all the potential targets available at any given moment. A complete assessment of computer assistance must include high workload conditions.

Signal Modification

Modern processing techniques permit changing the spectral or temporal pattern of the incoming data, to increase the perceptibility of potentially useful information. Modification techniques are based on a knowledge of basic psychoacoustics and the characteristics of potential targets. Two direct methods of modification are: changing the frequency range of the sound and filtering out "non-essential" frequency bands. The principle seems reasonable, but the usefulness of these methods of signal modification has not been clearly demonstrated.

Webster, Carpenter, and Woodhead (1968b) examined the effect of frequency doubling and high-pass filtering (together and separately) on the identification of sounds that differed only in timbre. They found no benefit (or detriment) with either type of modification. One limitation with this study was that the frequency doubling was done by speeding up a recording of the signals.

Since the whole session was pretaped, both the signal duration and the response time were halved. This made the task somewhat more difficult. A later study (Webster, Woodhead and Carpenter, 1969) examined the effect of frequency shifting by factors of 2, 4, or 8 on the classification of diesel engine sounds. In this study, signal and response duration was constant across all conditions. Once again, the effect of frequency shifting was negligible. They concluded that subjects were probably using the high-frequency formants, rather than the pattern among the harmonics, which should have been enhanced by frequency doubling, to identify the engine speed. A third study, (Webster, Woodhead, and Carpenter, 1970) examined the effect of frequency shifting on the recognition of vowel like, music like and meaningless (buzz) sounds. No benefit was found to presenting sounds, other than speech sounds, in the speech range. The meaningless sounds were recognized best at their natural fundamental (55 Hz), while the music-like sounds were least affected by frequency shifting.

The studies by Webster and his coworkers used simple, general processing techniques. The results might have been different if the modifications were tailored to the sound. Since these studies were completed, better tools have been developed for analysing the features a listener uses to identify a complex sound. Some of these techniques have been discussed above, in the section on feature analysis; others can be found in research on automatic speech recognition. Such analyses could be used to determine what features a listener uses, and processing techniques could then be applied to enhance them or increase their detectability. In addition, better processing techniques are available that allow independent manipulation of temporal and spectral characteristic of sounds.

Cueing

When an operator listens to a sound, he presumably compares it with stored representations of previously identified sounds. The better the set of stored representations the more accurate will be his classification of the new sound. Classification might be aided by providing labelled templates or cues for the operator to use as comparisons. Several studies have examined the usefulness of cues for enhancing detection and identification. Cues appear to be useful to the extent they provide information that is not otherwise available to the listener.

The use of cues to enhance detection has mainly been studied with simple sounds. Usually, the cue has the same characteristics as the potential signal and is presented either to the same ear as the potential signal or to the contralateral ear. When a cue is presented to the same ear, the detectability of a signal may be increased as much as 3 dB if the frequencies of the cue and signal are within a critical band. Unfortunately, a critical band is wide enough that several non-target signals could also be enhanced in a cluttered operational environment.

A contralateral cue has proved to be useful only if it provides precise information about the signal. For instance, Taylor and Forbes, (1969) found that a contralateral cue enhanced the detectability of a band-limited noise signal only when it provided moment to moment information about the waveform. Their cue was a replica of the masked signal. In the operational world, precise information about the form of the signal is not available. McFadden (1981) tried to get around the problem of not having precise information by presenting a cue whose frequency drifted across the frequency range in which a target was expected. Both signal and cue were harmonic complexes of the same character. No effect was found.

The detectability of a signal improves as the listener's knowledge of it and of the masking noise increases (Spiegel and Green, 1981). In the cueing studies, the listeners knew a great deal about the signal because it was repeated from trial to trial. It is not surprising that precise information about the signal was needed before detection improved further. In normal conditions, the listener's knowledge about a potential signal is poorer and less precise information may be useful. Gilliom (1976) found that observers could use a contralateral cue, presented before a signal, to reduce frequency uncertainty. The signal to be detected could be at any one of several frequencies. If a tone of the same frequency or a tonal complex with the centre tone the same frequency as the signal was presented to the contralateral ear before the signal was presented, detectability of the signal improved.

Cues may be used to aid target identification as well as detection. Mackie (1974) summarized a study that provided experienced operators with a tape of signal samples to use when classifying unknown targets. Classification performance was compared with and without this memory aid. The operators did significantly better with the aid. The greatest improvement was found with operators that did poorest initially. Possibly, the experienced operators' knowledge about the potential signals was good enough that the taped signals provided little additional information.

Further research is needed on aural templates as a classification aid. The study reported by Mackie suggests that templates are most useful for inexperienced operators. Possibly, a different type of aural template would be useful for experienced operators. What is the best form of template? Should whole signatures be used, or common features? How realistic should the target samples be?

Relevance to Aural Displays

Display enhancement is intended to increase the useful information available to the operator. Information can be increased by noise reduction or by providing alternate sources of information. In the studies reported here, the most successful methods were those that provided alternate sources of information. Efforts to reduce noise by filtering or by automatic processing have been less successful.

The ineffectiveness of methods such as an automatic alert may be as much a function of methodology as of intrinsic merit. In laboratory studies, the data rate is relatively low, and the operator has time to process all the incoming information. In the operational environment, the data rate is too high for the operator to process everything. Under such conditions, a systematic reduction of information may be superior to the operator's own *ad hoc* procedures. Preprocessing of stimuli should be tested under realistic workloads.

In contrast, the use of templates and alternate displays increases the information that the operator must process, which could cause problems in high workload conditions. Templates could have a further drawback. When the incoming sound could be one of a hundred targets instead of one of twenty, not only does the time required to check the templates increase, but so does the memory requirement. Care must be taken to ensure that the templates enhance the operator's capabilities. For instance, templates would probably be more effective if the operator used them to decide if the current sound was one of two similar targets than if he used them for initial classification.

CONCLUSIONS

The Current State

There is a substantial body of direct and indirect research on the detection and identification of complex nonspeech sounds. Psychoacoustic research suggests the ear is good at the detection of rapid shifts and the discrimination of subtle changes in frequency and amplitude. Psychoacoustic data also provide a source of knowledge for the development of models of peripheral processing. Such models are used to predict limits on the capability of a listener to identify a sound and to design methods for automatic preprocessing of sounds.

Research on pattern recognition and information processing suggests that the ear identifies complex aural sounds by extracting the underlying features from the sound and comparing the set of extracted features with those of known sounds. Recently, mathematical algorithms have been developed that may improve our capability to understand, or at least mimic, the aural processing done by the ear. The importance of knowledge driven or top-down processing has been recognized with the realization that the listener's capability to identify complex sounds in the real world is usually better than mechanical implementations have been able to achieve. Many different types of information may be used in the identification of a complex sound. Research emphasis in top-down processing has been on finding the kinds of knowledge that can influence perception.

Applied research on aural displays has emphasized training and enhancement. For enhancement, the most successful methods have been (i) bisensory displays that provide complementary information, (ii) the presentation of aural and written templates and (iii) the use of automatic pattern recognizers for informational and alerting purposes. Some of the techniques, such as the use of templates, have proven effective in training as well, but in general, extensive training still appears necessary. The development of computer aided learning systems may reduce the cost and increase the effectiveness of training programs.

What needs to be Done?

To make the most effective use of the ear, one must:

- determine precisely when aural displays should be used and what form they should take;
- minimize training time while maximizing the effectiveness of the training received;
- optimize the role of the machine so that it enhances the performance of the operator.

Possible methods for achieving the above goals appear to exist, but considerable research remains to be done. The work reviewed in this paper provides some direction for future research.

Future research on the form of aural displays should concentrate on determining how the aural display can be enhanced. Hitherto, aural displays for passive sonar have presented unprocessed returns. If it were possible to improve the perceptibility of the information available aurally, the amount of training and experience required should be reduced. To develop effective enhancements, more information is needed on the underlying structure of the sounds themselves. What about the target does the ear hear? With this information, it should be possible to modify the sound to enhance the relevant characteristics. Possible transforms include frequency shifting of low frequency sounds into the speech range; speeding up the rate at which a sound is played to make slow temporal fluctuations more audible, and filtering out irrelevant characteristics.

Another possible area for research is on the use of aural display for initial detection on multiple channel sonar systems. Currently, visual displays are used for detection (often in conjunction with an automatic alert system). With multiple channel sonars, information, summed or selected over frequency, is presented on a bearing-time-intensity display (BTI) or individual bearing are shown on multiple lofargrams. Generating and searching multiple lofargrams can be time consuming, and the signal to noise ratio is degraded by the summation involved in a BTI display. An aural display that presented the returns from each bearing in succession might result in better performance. If the signal frequencies are in the frequency and temporal region of the ear's competence, the ear is good at optimizing the bandwidth of its internal filters. It is also good at detecting broadband sounds and sounds which vary over time. A time compressed scanning aural display, (if time compression proved beneficial) would allow efficient and effective processing of a multiple channel sonar.

Future training studies should evaluate numerical methods for extracting the features used to identify complex sounds. Studies are also needed on how and to what extent situational information comes into the identification process. Previous research has shown that learning can be affected by manipulating the information presented to the listener about the targets. The information usually consists of descriptions of the characteristics of the individual targets, but contextual information also may be used for identification. Its role has

received much less attention.

With the continuing increase in usable computing power, future systems should provide the operator with more powerful tools for handling and analysing the incoming information. Studies are needed to ensure that these tools do help the operator. Research on optimizing the role of the machine should include studies on the value and form of automatic processing and on the use of templates and other informational aids. For instance, principal components analysis should be studied as a method of analysing incoming sounds. Research on templates should examine whether and when it is helpful to provide an operator with examples of potential targets or the dominant feature of relevant targets.

In all of the suggested studies, a prime concern of the researcher must be the applicability of the results to the operational environment. It is important that any potential display or training method not only improve performance but also improve efficiency. The amount of data to be processed on a passive sonar system is steadily increasing. Improved techniques are those that allow the operator to process information efficiently as well as effectively.

To determine what the best techniques are two basic questions must be answered: what role should the ear play in the processing of passive sonar information; and how can the best use be made of the ear in that role. Answering the first question involves analysing the tasks carried out by the sonar operator and determining the best sense or combination of senses to handle each task. For example — what type of initial detection display allows the operator to handle the most data, at the best signal to noise ratio with the least deterioration in performance over time. Possible solutions are automatic alerts, BTI, multiple lofargrams, a scanning aural display or some combination. The answer will depend on factors such as the capabilities of each sense, the characteristics of the targets to be processed, the available processing time per channel and the preprocessing capabilities.

Studies on display enhancement and training should provide answers to the question of how to make the best use of the ear. To do so, they must evaluate performance under realistic conditions. The best techniques will be those that produce both rapid and accurate classification either by increasing the discriminability of targets or by effective training of operators.

Conclusion

Aural displays should be used in passive sonar, but they should be developed so that they can be used more effectively than they have in the past. They provide a different and potentially important source of information that can complement the information provided by visual displays. What is needed is that the exact nature of that information be determined and techniques be developed to enhance its perceptibility.

Numerical methods such as principal components analysis and multidimensional scaling are available for analysing complex sounds. At the same time, research in psychoacoustics and information processing provides information on the basic capabilities of the human operator. The area where information is lacking is in the operational usefulness of the concepts and constructs developed in basic research. Future research on aural displays should attempt to incorporate the results of basic research to determine what role the ear should play and how to use the ear most effectively in that role.

REFERENCES

ANDREWS, F. A., & J. D. HOVATER. *Aural Detection By Passive Sonar.* SAO 71-16, Department of the Navy, Anti-Submarine Warfare Systems Project Office, Washington D.C. 20360, 1971.

ANNETT, J. *Sonar Recognition Training: An Investigation of Whole VS Part and Analytic VS Synthetic Procedures.* NAVTRADEVCEN 67-C-0105-1, Naval Training Device Center, Orlando Florida, 1971.

BABCOCK, H., & S. SUTTON. End Point of Lateralization for Dichotic Clicks. *J. Acoust. Soc. Amer., 39,* 87-102, 1966.

BLODGETT, H. C., W. A. WILBANKS, & L. A. JEFFRESS. Effect of Large Interaural Time Differences upon the Judgement of Sidedness. *J. Acoust. Soc. Amer., 28,* 639-643, 1956.

BREGMAN, A. S. The Formation of Auditory Streams. Chapter 4 in J. Requin (Ed.), *Attention and Performance VII.* Proceedings of the Seventh International Symposium on Attention and Performance, Senanque, France, August 1-6, 1976, Hillsdale, N. J.: Lawrence Erlbaum Associates, 1978.

BREGMAN, A. S., & J. CAMPBELL. Primary Auditory Stream Segregation and Perception of Order in Rapid Sequences of Tones. *J. Exp. Psychol., 89,* 244-249, 1971.

BREGMAN, A. S., & G. L. DANNENBRING. The Effect of Continuity on Auditory Stream Segregation. *Percept. Psychoph., 13,* 308-312, 1973.

COLE, R. A., & J. JAKIMIK. Understanding Speech: How Words are Heard. Chapter 3 in G. Underwood (Ed.), *Strategies of Information Processing.* London: Academic Press Inc. (London) Ltd., 1978.

CORCORAN, D. W. J., A. CARPENTER, J. C. WEBSTER, & M. M. WOODHEAD. Comparisons of Training Techniques for Complex Sound Identification. *J. Acoust. Soc. Amer., 44,* 157-167, 1968.

CORCORAN, D. W. J., J. L. DENNETT, & A. CARPENTER. Cooperation of Listener and Computer in a Recognition Task: Effects of Listener Reliability. *J. Acoust. Soc. Amer., 47,* 480-488, 1970.

CORCORAN, D. W. J., J. L. DENNETT, & A. CARPENTER. Cooperation of Listener and Computer in a Recognition Task. II. Effects of Computer Reliability and "Dependent" versus "Independent" Conditions. *J. Acoust. Soc. Amer., 52,* 1613-1619, 1972.

DURLACH, N. I., & H. S. COLBURN. Binaural Phenomena. Chapter 10 in E. C. Carterette, & M. P. Friedman (Eds.), *Handbook of Perception Volume IV: Hearing.* New York: Academic Press Inc., 1978.

FETH, L. L., & H. O'MALLEY. Two-Tone Auditory Spectral Resolution. *J. Acoust. Soc. Amer., 62,* 940-947, 1977.

GILLIOM, J. D., & W. M. MILLS. Information extraction from contralateral cues in the detection of signals of uncertain frequency. *J. Acoust. Soc. Amer., 59,* 1428-1433, 1976.

GREEN, D. M. Detection of Multiple Component Signals in Noise. *J. Acoust. Soc. Am., 30,* 904-911, 1958.

GREEN, D. M. Detection of Auditory Sinusoids of Uncertain Frequency. *J. Acoust. Soc. Am., 33,* 897-903, 1961.

GREEN, D. M., T. G. BIRDSALL, & W. P. TANNER. Signal Detection as a Function of Signal Intensity and Duration. *J. Acoust. Soc. Am., 29,* 523-531, 1957.

HALPERN, J. *Bisensory Information Processing: Discrimination as a Function of Cue Redundancy and Task Difficulty.* TR-6, Computer Image Corporation and University of Denver, Department of Psychology, 1971.

HARTLEY, L. R., R. G. OLSSON, & J. D. INGLEBY. *An Unbiassed Assessment of Simulated Computer Assistance in a Signal Detection Task.* OES-1/74, Royal Naval Personnel Research Committee, London, England, 1972.

HENNING, G. B. Frequency Discrimination of Random-Amplitude Tones. *J. Acoust. Soc. Amer., 39,* 236-339, 1966.

HENNING, G. B. Frequency Discrimination in Noise. *J. Acoust. Soc. Am., 41,* 774-777, 1967.

HENNING, G. B. A Comparison of the Effects of Signal Duration on Frequency and Amplitude Discrimination. In R. Plomp, & G. F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing.* Proceedings of the International Symposium on Frequency Analysis and Periodicity Detection in Hearing, Driebergen, The Netherlands, June 23-27, 1969, Leiden: A. W. Sijthoff, 1970.

HENNING, G. B., & S. L. BLEIWAS. Amplitude Discrimination in Noise. *J. Acoust. Soc. Am., 41,* 1385-1386(L), 1967.

HENNING, G. B., & J. PSOTKA. Effect of Duration on Amplitude Discrimination in Noise. *J. Acoust. Soc. Am., 45,* 1008-1013, 1969.

HOWARD, J. H. *The Psychophysical Structure of Complex Underwater Sounds.* ONR-76-2, Human Performance Laboratory, The Catholic University of America, Washington, D.C. 20064, 1976.

HOWARD, J. H. Psychophysical Structure of Eight Underwater Sounds. *J. Acoust. Soc. Am., 62,* 149-156, 1977.

HOWARD, J. H., & J. A. BALLAS. *Feature Selection in Auditory Perception.* ONR-78-5, Human Performance Laboratory, The Catholic University of America, Washington, D.C. 20064, 1978.

HOWARD, J. H., & J. A. BALLAS. *Sequential Structure and Context in the Classification of Nonspeech Transient Patterns.* ONR-80-14, Human Performance Laboratory, The Catholic University of America, Washington, D.C. 20064, 1980.

HOWARD, J. H., & J. A. BALLAS. Feature Selection in Auditory Perception. Chapter 11 in D. J. Getty, & J. H. Howard (Eds.), *Auditory and Visual Pattern Recognition.* Proceedings of Symposium on the Recognition and Classification of Aural (Non-speech) Signals, Cambridge, Mass., USA, June 28-30, 1978, Hillsdale, N. J.: Lawrence Erlbaum Associates, 1981.

HOWARD, J. H., J. A. BALLAS, & D. C. BURGY. *Feature Extraction and Decision Processes in the Classification of Amplitude Modulated Noise Patterns.* ONR-78-4, Human Performance Laboratory, Catholic University of America, Washington, D.C. 20064, 1978.

HOWARD, J. H., & D. C. BURGY. *Structure Preserving Transformations in the Comparison of Complex Steady-State Sounds.* ONR-78-6, Human Performance Laboratory, The Catholic University of America, Washington, D.C. 20064, 1978.

HOWARD, J. H., & E. B. SILVERMAN. *Structural Components in the Perception of Sixteen Complex Sounds.* ONR-75-1, Human Performance Laboratory, The Catholic University of America, Washington, D.C. 20064, 1975.

INGLEBY, J. D., P. HAMILTON, & A. K. COPEMAN. *Simulated Computer Assistance in a Signal-Detection Task (Its Effectiveness as a Function of Signal Probabilities, Cueing Time and Knowledge of Results).* OES-4/72, Royal Naval Personnel Research Committee, London, England, 1971.

JESTEADT, W., C. C. WIER, & D. M. GREEN. Intensity Discrimination as a Function of Frequency and Intensity Level. *J. Acoust. Soc. Amer., 61,* 169-184, 1977.

LOVELESS, N. E., J. BREBNER, & P. HAMILTON. Bisensory Presentation of Information. *Psych. Bull., 73,* 161-199, 1970.

MACKIE, R. R. *Research on Factors Influencing The Interpretation Of Sonar Signals.* 776-6, Human Factors Research, Incorporated, 6780 Cortona Drive Goleta, California 93017, 1974.

MACKIE, R. R., & A. HARABEDIAN. *A Study of Simulation Requirements for Sonar Operator Trainees.* NAVTRADEVCEN 1320-1, U. S. Naval Training Device Center, 1964.

MACKIE, R. R., C. D. WYLIE, R. R. RIDIHALGH, T. E. SHULTZ, & M. L. SELTZER. *Some Dimensions of Auditory Sonar Signal Perception and Their Relationships to Target Classification.* Tech. Rep. 2723-1, Human Factors Research Inc., Goleta, Calif. 93117, 1981.

MARSLEN-WILSON, W. D., & A. WELSH. Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychol., 10,* 29-63, 1978.

MARTIN, D. W. *Discrimination of Noise-Like Sounds Involving Multiple Interaction Features.* TM 78-171, The Pensylvania State University, Applied Research Laboratory, P. O. Box 30, State College, PA 16801, 1978.

MCFADDEN, S. M. *Detection of a Complex Tone in the Presence of a Contralateral Cue.* DCIEM Rep. No. 81-R-40, Defence and Civil Institute of Environmental Medicine, P.O. Box 2000, Downsview, Ontario, Canada, 1981.

MILLER, G. A., & S. ISARD. Some Perceptual Consequences of Linguistic Rules. *J. Verb. Learn. Verb. Beh., 2,* 217-228, 1963.

PATTERSON, R. D., & D. M. GREEN. Auditory Masking. In E. C. Carterette, & M. P. Friedman (Eds.), *Handbook of Perception Volume IV: Hearing.* New York: Academic Press Inc., 1978.

PINHEIRO, M. L., & H. TOBIN. Interaural Intensity Difference for Intracranial Lateralization. *J. Acoust. Soc. Amer., 46,* 1482-1487, 1969.

PLOMP, R. The Ear as a Frequency Analyzer. *J. Acoust. Soc. Am., 36,* 1628-1636, 1964.

PLOMP, R. Timbre as a Multidimensional Attribute of Complex Tones. In R. Plomp, & G. F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing.* Proceedings of the International Symposium on Frequency Analysis and Periodicity Detection in Hearing, Driebergen, The Netherlands, June 23-27, 1969, Leiden: A. W. Sijthoff, 1970.

REDDY, D. R. Speech Recognition by Machine: A Review. *Proceedings of the IEEE, 64,* 501-531, 1976.

SAYERS, B. MCA. Acoustic-Image Lateralization Judgments with Binaural Tones. *J. Acoust. Soc. Amer., 36,* 923-926, 1964.

SIDLEY, N. A., E. WINOGRAD, & E. W. BEDARF. Stimulus-Identification Overlap in Learning to Identify Complex Sounds. *J. Acoust. Soc. Amer., 38,* 11-13, 1965.

SIDLEY, N. A., E. WINOGRAD, & E. W. BEDARF. Learning to Identify Complex Sounds: Prompting Versus Confirmation. *J. Acoust. Soc. Amer., 38,* 1050-1052, 1965.

SMITH, M. J. Making the Most of Towed Arrays: The Human Factor. *J. Nav. Sci., 9,* 145-156, 1983.

SPIEGEL, M. F., & D. M. GREEN. Signal and Masker Uncertainty with Noise Maskers of Varying Duration, Bandwidth, and Center Frequency. *J. Acoust. Soc. Amer., 71,* 1204-1210, 1982.

STALLARD, J. M., & J. M. LESLIE. *Psychoacoustics and Passive Sonar Detection.* NOLTR 74-27, Naval Ordnance Laboratory, Silver Spring, Maryland 20910, 1974.

SWETS, J. A., S. H. MILLMAN, W. E. FLETCHER, & D. M. GREEN. Learning to Identify Nonverbal Sounds: An Application of a Computer as a Teaching Machine. *J. Acoust. Soc. Amer., 34,* 928-935, 1962.

TAYLOR, I. K., & M. M. TAYLOR. *The Psychology of Reading.* New York: Academic Press, 1983.

TAYLOR, M. M., & S. M. FORBES. Monaural Detection with Contralateral Cue (MDCC). I. Better than Energy Detector Performance by Human Observers. *J. Acoust. Soc. Amer., 46,* 1519-1526, 1969.

TAYLOR, M. M., S. M. SMITH, & M. E. TYTLA. Auditory Filter Shape by Subthreshold Summation. *J. Acoust. Soc. Amer., 57,* S76(A), 1975.

URICK, R. J. *Principles of Underwater Sound.* New York: McGraw-Hill Inc., 1975.

WATSON, C. S., & R. W. GENGEL. Signal Duration and Signal Frequency in Relation to Auditory Sensitivity. *J. Acoust. Soc. Amer., 46,* 989-997, 1969.

WATSON, C. S., & W. J. KELLY. The Role of Stmulus Uncertainty in the Discrimination of Auditory Patterns. Chapter 3 in D. J. Getty, & J. H. Howard (Eds.), *Auditory and Visual Pattern Recognition.* Hillsdale, New Jersey: Lawrence Erlbaum Assoc., 1981.

WEBSTER, J. C., A. CARPENTER, & M. M. WOODHEAD. Identifying Meaningless Tonal Complexes. *J. Acoust. Soc. Amer., 44,* 606-609, 1968.

WEBSTER, J. C., A. CARPENTER, & M. M. WOODHEAD. Identifying Meaningless Tonal Complexes II. *J. Aud. Res., 8,* 251-260, 1968.

WEBSTER, J. C., M. M. WOODHEAD, & A. CARPENTER. Identifying Diesel Engine Sounds. *J. Sound Vib., 9,* 241-246, 1969.

WEBSTER, J. C., M. M. WOODHEAD, & A. CARPENTER. Perceptual Constancies in Complex Sound Identification. *Br. J. Psychol., 61,* 481-489, 1970.

WEBSTER, J. C., M. M. WOODHEAD, & A. CARPENTER. Perceptual Confusion between Four-dimensional Sounds. *J. Acoust. Soc. Amer., 53,* 448-456, 1973.

WOODHEAD, M. M., A. CARPENTER, & J. C. WEBSTER. *Learning to Identify Sounds Using explicit Instructions.* RNPRC OES-74/5, Royal Naval Personnel Research Committee, London, England, 1973.

Defence and Civil Institute of Environmental Medicine
1133 Sheppard Avenue West, Post Office Box 2000
Downsview, Ontario, Canada, Telephone (416)633-4240